图书情报工作
第 65 卷 第 23 期　2021 年 12 月

网络舆情观点团簇演化等级测度与实证研究

■ 闫璐　杨刚　赵江元
吉林大学管理学院　长春 130022

摘　要：[目的/意义]提出和构建网络舆情观点团簇演化等级，以描述网络舆情受众的群体性观点的状态随时间与事态变化的演化程度，对于网络舆情导控与精准引导具有重要的理论及实践意义。[方法/过程]基于 LDA 与 CNN 神经网络构建网络舆情观点团簇演化等级测度模型，并以"翟天临知网事件"为实验对象，验证演化等级这一指标的有效性。[结果/结论]网络舆情观点团簇演化等级能够很好地体现网络热点事件群体观点状态的演化，在展现 3 个维度的属性数值同时也能反映观点团簇较前一时间节点状态的演化程度，提出的观点团簇演化等级测度结果精准地体现事件观点的各个演化高峰，为有关部门对网络舆情群体观点的靶向引导提供新的指导方向。

关键词：网络舆情　观点团簇　观点演化　演化等级　测度模型
分类号：G203
DOI：10.13266/j.issn.0252-3116.2021.23.012

1　引言

在当前互联网技术快速革新的今天，新媒体的蓬勃发展推动了网络用户数量的高速增长，形成了当前复杂多变的网络舆情生态环境。网络舆情个体观点是基于个体的立场、观念、认知水平等思维模式所形成的，具有极强的主观性，在个体观点形成的过程中受其他观点的影响，个体之间的交互和相互影响与演化促成了关于某类网络热点事件的几种主流观点产生，这种主流群体观点的抽象概念化可以认为是由多个具有相同或相似立场的个体观点形成的团簇结构，即观点团簇。观点团簇的演化则是观点团簇的情感倾向、团簇规模、传播范围等演化属性随事件发展或时间推移所产生的变化，当观点团簇在后续发展过程中其属性发生变化则表明该观点团簇内的网民对事件的态度发生了改变。对网络舆情观点团簇演化的感知有利于舆情监管部门对网络舆情发展情况的把握，及时发现舆情事件的观点变化能够规避潜在舆情风险以及消除舆情危机。

笔者基于 LDA（latent dirichlet allocation）模型与 CNN（convolutional neural networks）神经网络构建网络舆情观点团簇演化等级测度模型，网络舆情观点团簇演化等级作为描述网络舆情受众的群体性观点演化程度的一种指标，其演化等级的高低体现了观点团簇的演化程度，该指标能够较好地体现网络舆情随时间维度与事态发展的演化状态，包括网络舆情观点团簇的情感变化、影响能力、观点体量等信息，能够为网络舆情管控主体针对性的精准引导舆情风向与管控舆情发展提供参考依据。笔者在理论层面明确观点团簇概念与其演化属性，在实践层面构建网络舆情观点团簇演化等级指标与其测度模型，并利用"翟天临知网事件"作为研究实例，验证观点团簇演化等级与测度模型的泛用性与准确性。

2　相关研究

2.1　网络舆情观点挖掘

观点挖掘主要是对观点的情感倾向性与语义信息进行提取，有研究者利用 K-最邻近分类器与朴素贝叶斯分类器结合生物信号检测的方法获取用户的观点[1]。也有学者利用 word2vec 将观点文本向量化后输入 SVM（Support vector machines）与 LSTM（Long short-term memory）模型中，训练出能够识别新观点的观点挖掘模型[2]。基于深度学习的观点挖掘方法则是目前研究中使用的主流方法，有研究搭建带有 Dropout 机制的多层嵌入 CNN 模型，增强模型的局部语义特征识别能力，能较好地挖掘带有某种观点的文本[3]，也有研究者

作者简介：闫璐，博士研究生，E-mail：623716346@qq.com；杨刚，教授，博士，博士生导师；赵江元，博士研究生。
收稿日期：2021-06-20　修回日期：2021-09-23　本文起止页码：106-115　本文责任编辑：徐健

通过结合词向量与多尺度卷积神经网络对网络舆情观点情感倾向进行分类,将3种尺度卷积单元融合为一维向量,在实际观点挖掘任务中有着良好的表现[4]。

综上所述,在网络舆情观点挖掘的研究上,研究者主要利用各类分类器对网络舆情的群体观点进行挖掘,但从群体观点的属性状态角度分析其随时间与事态发展而产生变化的研究相对较少,笔者结合语义角度、情感类型、传播能力等特征综合挖掘网络舆情的观点有利于从多元角度分析当前网络舆情观点的演化状态,在该领域展开相关研究能够为网络舆情观点演化的引导与管控提供实践上的指导。

2.2 网络舆情观点演化

目前,对于网络舆情观点演化的研究主要从传播的角度表现观点的演化情况。有研究者从网络舆情用户信息及文本内容视角出发,构建不同维度的网络舆情主题图谱,结合主题图谱对网络舆情进行特征演化及可视化分析,以表示网络舆情观点的演化特征[5];有学者提出和构建网络舆情衍进指数,以文本聚类结果和文本聚类有效性为依据,提出网络舆情衍进的判别标准,以描述网络舆情发展过程中主流观点演化以及新观点产生的过程[6];有研究基于社会网络模型构建相邻节点之间的连续观点交互模型,以表示观点在传播过程中的演化状态[7];目前也有研究基于传染病模型分析网络舆情观点传播的时间演化特征,从而明确网络舆情的爆发节点,实现无监督预警[8]。

综上所述,目前国内外学者对网络舆情观点的演化主要集中于利用社会网络模型或传播模型对观点团簇的演化状态进行表示,并没有通过量化方式以数据的形式表现网络舆情热点事件中广大网络舆情受众的主流观点的数值特征,网络舆情观点团簇的演化量化方向尚存在研究空间。

3 网络舆情观点团簇演化等级测度模型构建

3.1 观点团簇演化等级测度流程

笔者提出的观点团簇演化等级测度的实现思路为:首先对网络舆情观点数据进行观点团簇划分,并对各个观点团簇的情感强度、团簇规模与传播范围进行量化,作为观点团簇的演化属性。然后,根据观点团簇中关键词的语义相似性构建观点团簇演化链,以此表示某一主题的观点团簇随时间延续的状态。最后,根据演化链上观点团簇演化属性在各个时间节点的增长情况得出观点团簇演化等级,作为观点团簇演化高峰的研判依据。

观点团簇演化等级测度的关键在于观点团簇的划分上,首先需要确定网络舆情观点的情感分类,然后再对正负情感观点进行观点团簇划分,这样做的目的是解决LDA划分观点团簇后其中包含正负情感观点数量接近、不能很好地反映观点团簇情感倾向性的问题。因此,观点团簇的划分过程分为3步:首先利用卷积神经网络与情感词典相结合的方法,对网络舆情观点数据进行情感分类同时计算情感强度,然后按照固定时间窗对正负情感语料进行切片,笔者以天为单位对网络舆情数据进行切片,然后利用困惑度指标确定每个时间片中正负情感观点团簇的最佳划分数量K,最后利用LDA模型分别对正负情感语料进行观点团簇划分。

观点团簇演化等级测度模型的构建流程为:①确定观点极性。划分正向情感观点与负向情感观点,并计算情感强度。②划分观点团簇。利用LDA模型,根据困惑度指标确定最佳观点团簇分类数,划分正向观点团簇与负向观点团簇。③观点团簇演化属性量化。对观点团簇的传播范围、情感强度、团簇规模3种演化态势属性进行量化。④构建观点团簇演化链。根据观点团簇中观点的文本语义相似度确定前后时刻观点团簇的演化关系链条。⑤观点团簇演化等级计算,对比同一演化链中前后时间节点观点团簇的演化属性的变化情况,得出观点团簇的演化等级。

3.2 观点团簇情感强度测度

笔者利用卷积神经网络(CNN)对网络舆情数据进行情感分类,同时结合情感词典计算每条观点的情感强度。CNN能极为准确地对网络舆情观点的情感进行分类,而情感词典方法则能以数值体现情感的强度,将二者结合即可得到观点的情感强度。卷积神经网络(CNN)情感分类的方法与CNN处理图像的方式类似,通过卷积层提取特征,然后通过池化层减少神经元数量,最后通过全连接层作为分类器输出概率。笔者首先根据词频构建词与频次的字典,词频越高的词排序越靠前,保留前一万个词以加快训练速度。此时已将中文词汇转换为模型可读的数据类型,然后利用卷积层分别以三个词、四个词、五个词的移动步长读取句子作为卷积核,已经能够完美地呈现句子的语义内涵[4]。利用交叉熵函数作为损失函数计算语料训练中的损失。最后,将多种卷积核提取的特征向量展开并连接在一起,并加全连接层输出类别。根据损失函数与准

确率的变化情况对卷积神经网络参数进行调整，CNN 模型参数如表 1 所示：

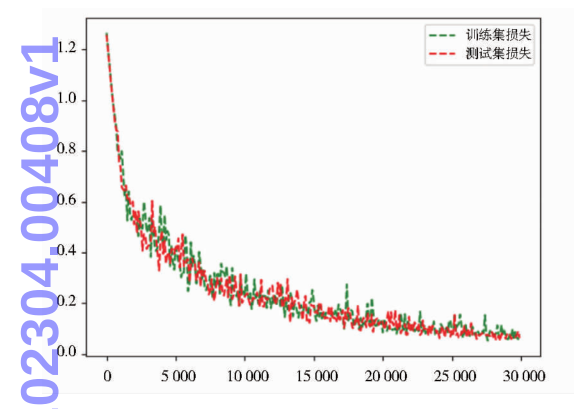
表 1 CNN 模型参数

参数	释义
vocab_size = 10 000	保留频次前一万的词
max_seq_num = 256	每个句子最多词数量
num_dimensions = 100	词向量维度
batch_size = 64	batch 移动的步长
Filter_sizes = [3,4,5]	三种卷积核尺寸
num_filters = 32	卷积核数目
num_classes = 2	输出类别
Iterations = 10000	迭代次数
Dropout = 0.5	Dropout 保留比例
Learn_rate	学习率为 0.001

在确定模型参数后，情感分类损失与准确率见图 1。

从图 1 可以看出，CNN 模型在使用表 1 参数后的损失函数在 3 万次循环后已经收敛，且分类准确率达到了较高的水平。

在微博舆情观点情感强度确认后，利用 BosonNLP 情感词典[9]，结合停用词、否定词、程度副词词典的方式对微博舆情观点文本内容进行情感强度计算，利用 python 的 jieba 分词包对测试集语料进行分词，并去停用词，将切分的词与情感词典进行匹配，最后得出含有分数的情感值。



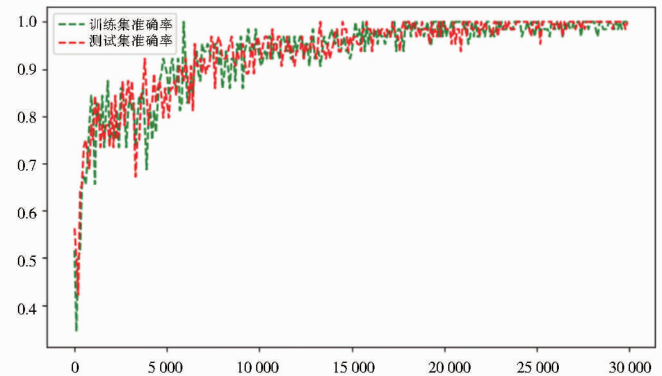


图 1 CNN 模型测试集与训练集损失变化

3.3 观点团簇规模测度

在观点团簇规模测度上，笔者选用 LDA 模型作为观点团簇划分模型，LDA 作为一种无监督模型，虽然在短文本分类效果上不如长文本，但能够较好地应对大规模网络舆情信息无法人工标注类型的问题，根据文档集中的每篇文档按照概率分布的形式给出分类结果。网络舆情观点团簇的主题即可认为是团簇内观点共同的观点反应，基于 LDA 的网络舆情观点团簇划分的思想是假设网络舆情观点本体中的每个词都以一定概率描述某一观点团簇主题，而观点团簇主题是由一组特征词的概率分布来描述，不同主题可能包含词汇相同但属于团簇概率不同的特征词，因此每类观点团簇主题的高频词集合则可视作为该类话题潜在语义的体现[10]。在得到最终的观点团簇主题概率分布，迭代完成后输出观点本体 - 团簇主题矩阵和团簇主题 - 词的概率矩阵。根据词属于主题的概率将词插入团簇主题列表中构成主题的特征词集合即代表观点主体对微博事件的观点的表述。

在观点团簇划分过程中，最优观点团簇数量将影响观点团簇划分的精确与否，利用 LDA 模型进行观点团簇划分的一个问题是单位时间内观点团簇数量 K 的值，即语料库的最优分类数确定问题，因此利用 LDA 主题模型分类需要确定分类个数，笔者采用较为通用的困惑度指标(Perplexity)确定每个时间段中的最佳观点团簇个数，以此来体现分类的可信度。困惑度用于度量概率分布或概率模型预测样本的好坏程度，通过对比两个概率分布或概率模型在预测样本上的优劣来选取最优模型，困惑度在评价聚类分类算法的性能上有极好的效果，因此可以通过对比困惑度来选取 LDA 的最优观点团簇数。困惑度针对不同模型从概率分布困惑度、概率模型困惑度以及分词困惑度 3 种方法计算困惑度，针对自然语言处理模型通常选用 Perplexity per word，即分词困惑度方法进行困惑度计算[11]。在测试集 Dt 上，困惑度表达式如公式(1)所示：

$$Perlexity(D_t) = \exp\left\{\frac{\sum_{d=1}^{M} \log p(w_d | M)}{\sum_{d=1}^{M} N_d}\right\} \quad 公式(1)$$

其中，M 是指训练好的模型参数，在 LDA 模型中为 theta 和 phi，即观点 - 观点团簇矩阵与观点 - 特征

词词矩阵，N_d 为观点 d 中单词数量，w_d 为测试集 D_t 中观点 d 的词向量形式。由公式(1)可知，中括号内分子项为测试集中微博 d 属于模型上的概率，其值越大代表困惑度越小，说明模型性能越高。

在对观点团簇最佳团簇数量求解时需要自动识别最佳观点团簇数量，但离散点组成的曲线无法直接提取最佳团簇数量，因此笔者采用曲线拟合的方式，将离散点组成的函数拟合成为连续函数并求其二阶导数，利用函数曲线二阶导数为零则为函数拐点的性质，对困惑度拟合曲线第一个最低点作为网络舆情观点团簇规模划分的最佳数目。网络舆情观点团簇困惑度曲线与拟合曲线如图 2 所示：

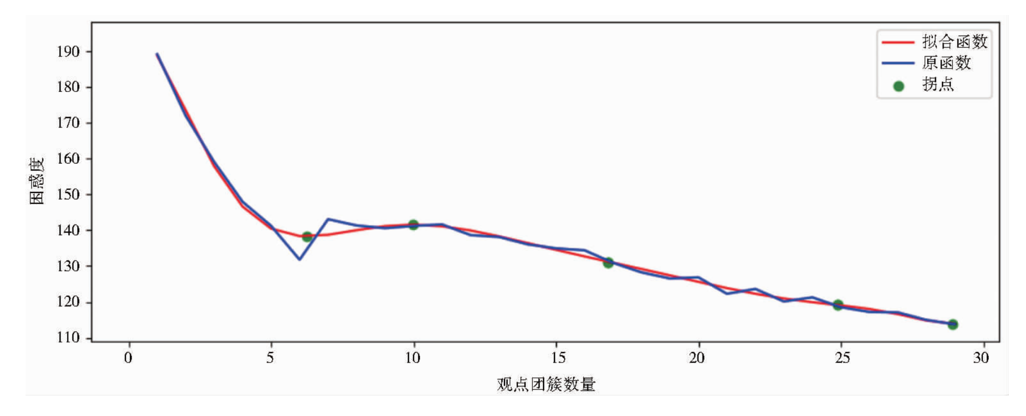


图 2　网络舆情观点团簇困惑度曲线与拟合曲线示意图

在确定好每个时间切片中网络舆情观点的最佳观点团簇数量后，利用 LDA 对每个时间段内的正负情感观点进行观点团簇划分，同时可以得到观点团簇的规模属性数据。

3.4　观点团簇传播范围测度

传播范围的计算思路为：分别计算观点团簇的用户传播能力与观点传播能力，然后将用户传播能力与观点传播能力结合得到观点团簇的传播范围。用户传播能力以网络舆情用户的账号互动属性即该用户的关注数量、粉丝数量、微博数量的总量表示，各指标取值与其权重和该用户的认证系数相结合即得到用户传播能力，在计算观点团簇内各个用户传播能力后即得到观点团簇的用户传播能力，因此观点团簇用户传播能力 C_{user} 的计算如公式(2)所示：

$$C_{user} = index_{rz} \sum_{i=1}^{N} n_{gz}w_{gz} + n_{fs}w_{fs} + n_{wb}w_{wb} \qquad 公式(2)$$

其中，$index_{rz}$ 为观点团簇内用户的认证系数，N 为观点团簇内用户数量，n_{gz}、n_{fs}、n_{bw} 分别为观点团簇内用户关注数量、观点团簇内用户粉丝数量、观点团簇内用户博文数量，w_{gz}、w_{fs}、w_{bw} 为指标权重。

观点传播能力与用户传播能力计算相同，以网络舆情观点的互动属性即该观点获得的点赞数量、转发数量、评论数量的总量表示观点传播能力，结合其指标权重与观点的数据类型系数求出网络舆情观点的传播能力，观点团簇的观点传播能力即为各观点累加求得，因此网络舆情观点团簇的观点传播能力 $C_{opinion}$ 的计算如公式(3)所示：

$$C_{opinion} = index_{type} \sum_{i=1}^{M} n_{dz}w_{dz} + n_{zf}w_{zf} + n_{pl}w_{pl} \qquad 公式(3)$$

其中，$index_{type}$ 为观点团簇内观点数据类型加成，M 为观点团簇内观点数量，n_{pl}、n_{dz}、n_{zf} 分别为观点团簇内观点评论数量、观点团簇内观点点赞数量、观点团簇内观点转发数量，w_{pl}、w_{dz}、w_{zf} 分别为各指标权重。

网络舆情观点团簇的传播能力则是由观点团簇用户信息量与网络舆情观点信息量相加得出。网络舆情观点团簇信息量 $C_{cluster}$ 的计算如公式(4)所示：

$$C_{cluster} = C_{user} + C_{opinion} \qquad 公式(4)$$

3.5　观点团簇演化链构建

根据观点动力学理论，个体的观点形成受其他个体观点的影响，且其观点情感倾向于个体所认同的观点[12]。观点团簇演化链可以认为在时间维度上，后续时间节点的观点团簇是受到前序节点观点团簇影响所产生的，二者在语义内涵上具有较高的一致性，因此当前后时间点上的两个观点团簇具有最高的语义相似度时，则可以认为二者在时间维度上是演化关系，后一时间节点的观点团簇是前一时间节点观点团簇在时间维度上的延续。观点团簇演化链构建流程见图 3。

笔者利用 Textrank 算法将观点团簇文本内容的分词进行排序，选取前 500 个词作为相似度对比依据，再利用 TFIDF 方法计算前后时间节点语料库中词的 TFIDF 值，将 TFIDF 值与观点团簇类别构成观点团簇的向量矩阵，通过余弦值相似度方法将两个时间节点中的观点团簇进行两两比较，利用余弦值相似度方法进行比对，能够得到前后两个时间节点的观点团簇相

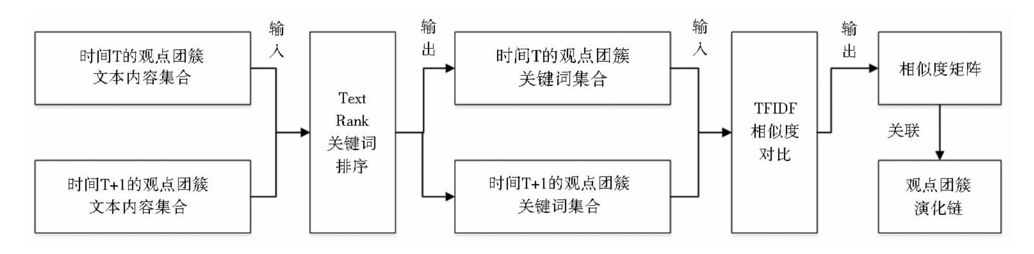


图 3 观点团簇演化链构建流程

似度矩阵,矩阵中数值表示后一时间节点中每个观点团簇与前一时间节点中每个观点团簇的语义相似度,通过提取相似度最高的两个观点团簇进行观点团簇演化链的构建。该方法能消除大量无用词对观点团簇语义相似度对比的干扰,同时对排名靠前的词进行相似度比对也更能体现观点团簇彼此之间的关联性。

首先需要将每个观点团簇中的网络舆情观点进行合并,将整合后的语料作为网络舆情观点团簇文本内容,则每类观点团簇以一个长文本表示,该长文本是该观点团簇内所有网络舆情观点汇总。某个事件的生命周期为 T 天,则 $T=\{1,2,3,\cdots,t\}$,时间节点第 t 天中含有 n 个网络舆情观点团簇,对观点团簇文本进行分词可得到第 n 个观点团簇的关键词语料集合为 $W_n=\{w_1,w_2,w_3,\cdots,w_i\}$,前一时间节点第 t-1 天中含有 m 个网络舆情观点团簇,第 m 个观点团簇的关键词语料集合为 $W_m=\{w_1,w_2,w_3,\cdots,w_j\}$,利用 Textrank 算法对所有观点团簇关键词语料进行重新排序,并选取前 500 个词,如果语料中不足 500 词则用空格填充。利用 TFIDF 方法计算前后时间节点中每个观点团簇语料库中词在该语料库中的 TFIDF 值。

TFIDF 值可以表征一个词对语料库中其中一份文档的重要程度,词的重要性随着其在文档中出现次数而增加,但同时会随着其在语料库中出现频率增加而降低[13]。在本文中经过排序后的观点团簇关键词的 TFIDF 值将作为观点团簇的向量表示。TFIDF 值为 TF 值与 IDF 值的乘积,词频(term frequency,TF)指某一词在该文档中出现的次数,通常对该值进行归一化处理以防止长文本干扰。

TFIDF 值体现了词代表某一观点团簇语义内容的重要程度,每个观点团簇选取最能代表团簇内涵的前 500 个关键词并计算器 TFIDF 值,将时间节点 t 中各个观点团簇与其所含关键词的 TFIDF 值关联构成了该时间节点的观点团簇向量,向量中每个值表示观点团簇中关键词的 TFIDF 值,500 个词则表示向量共 500 维,将观点团簇向量化后便可根据余弦相似度算法计算两个观点团簇之间的相似度。观点团簇的余弦值相似度计算如公式(5)所示:

$$cos_{similar}=\frac{\sum_{i=1}^{k}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{k}(x_i)^2}\times\sqrt{\sum_{i=1}^{k}(y_i)^2}} \quad \text{公式(5)}$$

其中,k 为向量维数,x_i、y_i 为前一时间节点与后一时间节点第 i 维向量值,余弦值数值越接近 1 则说明两个观点团簇向量余弦夹角越小,则相似度越高。将前后两个时间节点的向量进行相似度计算后可以得到 $n\times m$ 维的相似度矩阵,如公式(6)所示:

$$Matrix_{similar}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{bmatrix} \quad \text{公式(6)}$$

其中,n 为后一时间节点中观点团簇数量,m 为前一时间节点中观点团簇数量,在矩阵中列表示后一天第 n 类观点团簇与前一天所有观点团簇的相似度值,矩阵中每列的最大值的行数即表示后一时间节点观点团簇是前一时间节点中某个观点团簇的演化。将事件中所有时间节点进行观点团簇相似度对比后即可构建观点团簇的演化链,并根据观点团簇的演化链提取该演化链中观点团簇的演化态势属性。根据现实情况来看,前一天的观点团簇会存在多个演化链的情况,即后续时间节点中多个观点团簇都是受到前一团簇的影响形成的,因此利用相似度矩阵构建演化链时会存在后一时间节点的多个观点团簇与前一时间节点中某个观点团簇相似度皆最大的情况,属于演化中的分裂现象。

3.6 观点团簇的演化等级构建

在构建观点团簇演化链之后,便可对观点团簇演化链上的观点团簇演化属性分值进行计算。通过对已有事件的观点团簇演化情况进行分析得知,传播范围、情感强度以及团簇规模在增幅小于 100% 时,其观点团簇内涵、观点状态等属性基本无变化,因此其演化性较低;而增幅在 100% 至 200% 之间时,前后观点团簇具有较为明显的演化,演化性凸现;而当增幅大于

闫璐，杨刚，赵江元. 网络舆情观点团簇演化等级测度与实证研究[J]. 图书情报工作,2021,65(23):106－115.

200% 时观点团簇的演化性极高,其多数观点内涵发生了变化[14]。笔者对观点团簇演化属性分值用分段函数表示,观点团簇的传播范围、情感强度以及团簇规模的分值计算如下:

传播范围演化属性分数 Lc_t 如公式(7)所示:

$$L_{C_t}=\begin{cases}1 & if\ (0\leqslant\frac{C_t-C_{t-1}}{C_{t-1}}<1)\\ 2 & if\ (1\leqslant\frac{C_t-C_{t-1}}{C_{t-1}}<2)\\ 3 & if\ (2\leqslant\frac{C_t-C_{t-1}}{C_{t-1}})\end{cases} \qquad 公式(7)$$

其中,C_t 是演化链上时间节点 t 的观点团簇传播范围数值,C_{t-1}是前一时间节点 t－1 上观点团簇传播范围数值。

情感强度演化属性分数 Ls_t 的计算如公式(8)所示:

$$L_{S_t}=\begin{cases}1 & if\ (0\leqslant\frac{S_t-S_{t-1}}{S_{t-1}}<1)\\ 2 & if\ (1\leqslant\frac{S_t-S_{t-1}}{S_{t-1}}<2)\\ 3 & if\ (2\leqslant\frac{S_t-S_{t-1}}{S_{t-1}})\end{cases} \qquad 公式(8)$$

其中,S_t 是演化链上时间节点 t 的观点团簇情感强度数值,S_{t-1}是前一时间节点 t－1 上观点团簇情感强度数值。

团簇规模演化属性分数 Lo_t 的计算如公式(9)所示:

$$L_{O_t}=\begin{cases}1 & if\ (0\leqslant\frac{O_t-O_{t-1}}{O_{t-1}}<1)\\ 2 & if\ (1\leqslant\frac{O_t-O_{t-1}}{O_{t-1}}<2)\\ 3 & if\ (2\leqslant\frac{O_t-O_{t-1}}{O_{t-1}})\end{cases} \qquad 公式(9)$$

其中,O_t 是演化链上时间节点 的观点团簇规模数值,O_{t-1}是前一时间节点 t－1 上观点团簇规模数值。

笔者将网络舆情观点团簇的演化等级分为 14 个等级,每个级别根据观点团簇的传播范围、情感强度、团簇规模 3 个演化属性分数进行进一步划分,观点团簇演化属性分数 Lc_t、Ls_t、Lo_t 的取值分别通过公式(7)、公式(8)、公式(9)计算得出,计算观点团簇演化链上每个时间节点与后一时间节点上观点团簇演化属性分数,并与观点团簇演化等级分型表进行匹配,得到观点团簇的演化等级。网络舆情观点团簇演化等级分型如表 2 所示:

表 2 网络舆情观点团簇演化等级分型

演化等级	演化属性分数			网络舆情观点团簇演化状态表征
	传播范围	情感强度	团簇规模	
1 级	3	3	3	观点团簇传播范围、情感强度、团簇规模等方面全面演化,影响范围极广,极易形成网络舆情危机,其演化态势等级属于最高级别
2 级	3	2	3	
3 级	3	3	2	
4 级	2	3	3	
5 级	3	2	2	观点团簇的传播范围、情感强度与团簇规模增长适中,表明有部分的具有极强影响力的观点在较广的范围传播,存在成为舆情危机的潜在风险
	3	1	3	
6 级	2	2	3	
	1	3	3	
7 级	2	3	2	
	3	3	1	
8 级	2	2	2	
9 级	3	1	2	观点团簇传播范围、情感强度与团簇规模增长幅度较小,表明观点团簇处于较为活跃的演化状态,需要加强后续观察
	3	2	1	
10 级	2	1	3	
	1	2	3	
11 级	2	3	1	
	1	3	2	
12 级	3	1	1	
	1	1	3	
	1	3	1	
13 级	2	1	2	观点团簇的传播范围、情感强度与团簇规模变动幅度有限,属于观点团簇演化中的阶段性波动
	2	2	1	
	1	2	2	
14 级	2	1	1	
	1	1	2	
	1	2	1	
	1	1	1	

4 实证研究

4.1 数据源选择与采集

为验证网络舆情观点团簇演化等级的适用性,笔者以“翟天临知网事件”作为实证对象。在该事件中,第一次热度达到峰值是由于教育部出面回应该事件并展开调查,而第二次热度上升则是北京电影学院发布调查结果并取消翟天临博士学位,随后热度持续走低进入蔓延期。该事件的两次热度高潮都是由于事态出现转折,官方出面回应将事件热度拉向顶峰。而网友对于该事件的讨论多集中在对翟天临博士论文涉嫌抄袭、北京电影学院的调查结果以及教育部回应等方面。2 月 9 日至 2 月 15 日“翟天临知网事件”关键时间节点见图 4。

利用爬虫工具以“翟天临”为关键字从微博平台中抓取时间跨度为 2019 年 2 月 9 日至 2019 年 4 月 9 日

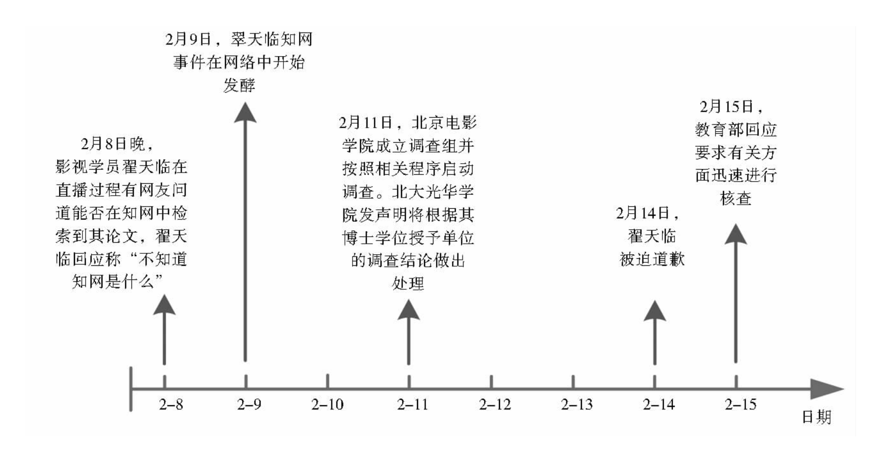


图 4　2 月 9 日至 2 月 15 日"翟天临知网事件"关键时间节点

的相关网络舆情数据并进行数据清洗，包括去除纯表情或符号博文、利用正则表达式消除相同样式前缀与后缀、去除同一博主重复发文等，得到共计 27 433 条有效数据，"翟天临知网事件"的观点数量随时间变化关系如图 5 所示：

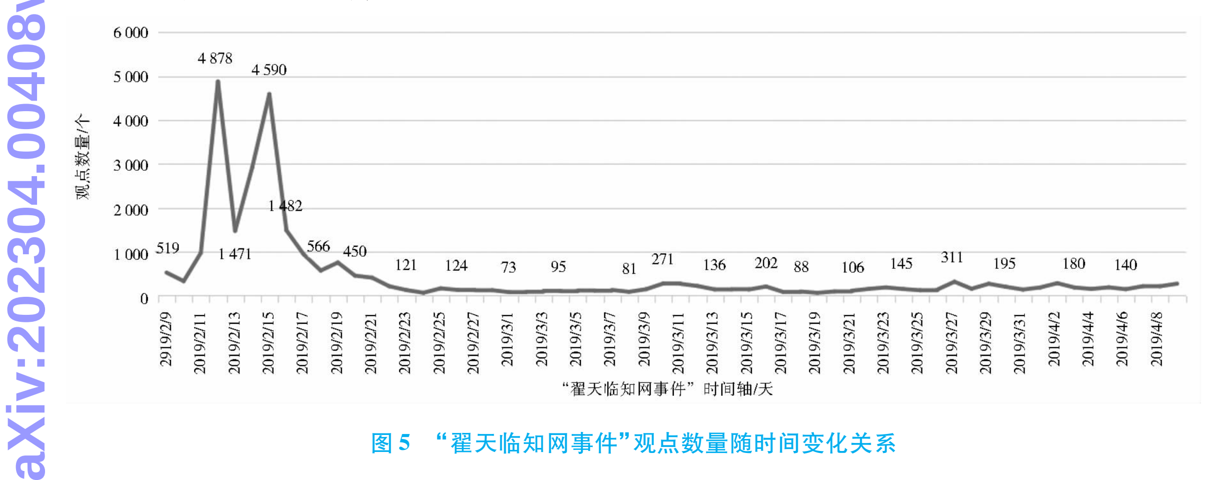


图 5　"翟天临知网事件"观点数量随时间变化关系

从图 5 中可以看出，"翟天临知网事件"中的微博舆情观点数量在 2 月 12 日与 2 月 15 达到高峰，随后快速下降，观点数量变化趋势与事件热度发展趋势基本相同，表明抓取的数据能很好地描述该事件发展情况。

4.2　数据处理与分析

（1）观点团簇演化属性测度。首先需要根据该事件中观点的情感强度将事件微博数据划分为正向情感观点与负向情感观点，经过情感分类后得到正向情感观点 5 153 个，负向情感观点 22 280 个，可见在该事件中，对于演员翟天临的博士论文抄袭广大网络舆情用户是持有负面情感。随后计算观点团簇的演化属性数值，由于事件跨越周期较长，因此笔者选取事件开始日期 2 月 9 日与最后一次热度高峰 2 月 15 日这两个关键时间节点展示网络舆情观点团簇的演化属性测度数据，2 月 9 日"翟天临知网事件"正负向网络舆情观点团簇属性信息如表 3、表 4 所示：

表 3　2 月 9 日正向观点团簇演化属性数值与语义内涵

团簇类别	传播范围	情感强度	团簇规模	语义内涵
1	1 254	190	16	大量明星参加新年晚会活动
2	12 209	242	24	对翟天临新作品《原生之罪》第二部的期待
3	4 069	201	17	关于美食的做法与功效
4	22 475	205	18	翟天临学术问题的报道
5	3 120	129	17	翟天临论文查重问题的报道
6	9 717	183	22	粉丝对于翟天临的赞美
7	1 680	139	19	旅行与文艺类观点

由表 2 与表 3 中可知，在 2 月 9 日正负网络舆情观点团簇数量皆为 7，由于事件正处于发酵前期，部分网友以及翟天临粉丝并不知晓翟天临事件的发生，因此存在几类观点团簇是对翟天临作品期待、对其演技认可或其参加活动的内容，也存在部分观点团簇是发表在翟天临超话板块的广告类内容，考虑到该部分"噪音"同样具有传播影响能力，故本文予以保留。正向观点团簇的各项属性数值皆较低，观点团簇规模普遍较

表 4　2 月 9 日负向观点团簇演化属性数值与语义内涵

团簇类别	传播范围	情感强度	团簇规模	语义内涵
1	104 649	-227	63	翟天临人设崩塌,学术造假
2	766 095	-328	61	对翟天临的讽刺,要求其做出回应
3	34 876	-197	54	对翟天临演技的认可但对其行为的批判
4	893 032	-319	59	批判翟天临名不副实,与博士学霸人设不符
5	1 015 614	-306	48	批判翟天临工作室意图降低事件热搜
6	415 996	-225	46	希望翟天临的学校给出回应
7	5 589	-198	55	借翟天临事件对真正学术研究的阐述

小。而负面观点团簇的属性数值则较大,最为明显的为负向第五类观点团簇,该团簇传播范围最大但团簇规模较小,表明此时已经有较多观点领袖参与事件讨论并提出负向观点,而根据语义内涵可知此时已有观点团簇指向翟天临学术论文抄袭、翟天临演技尚可但无德、希望翟天临给出解释等方面。

在 2 月 15 日,翟天临知网事件热度达到顶峰,该时间节点的正负向观点团簇属性信息如表 5、表 6 所示:

表 5　2 月 15 日负向观点团簇演化属性数值与语义内涵

团簇类别	传播范围	情感强度	团簇规模	语义内涵
1	44 349	2 132	201	生活文艺类观点
2	200 248	2 350	166	影视作品类观点
3	687 440	1 862	159	女性护肤品与化妆品广告类观点
4	148 093	2 186	157	建筑、旅游类观点
5	121 450	2 216	141	减肥与减肥食品类观点
6	52 813	1 999	161	家居装修类观点

表 6　2 月 15 日负向观点团簇演化属性数值与语义内涵

团簇类别	传播范围	情感强度	团簇规模	语义内涵
1	1 803 361	-4 090	700	网友批判娱乐圈教育问题
2	1 007 101	-2 802	561	对学术腐败问题的讨论
3	5 408 158	-3 398	534	北大与北电发表声明要彻查翟天临学术问题
4	8 011 850	-6 552	941	网友对学术公平性的讨论
5	40 682 983	-7 289	684	教育部回应翟天临事件的相关内容

由表 4 与表 5 可以得知,在舆论热度达到顶峰时各类观点团簇的演化属性数值极高,团簇规模与信息量等属性的数值较事件开始日期 2 月 9 日增长数十倍,表明此时该事件已经得到了广泛关注,通过观点团簇关键词总结的观点团簇语义内涵可以看出,此时正向观点团簇已经几乎没有微博舆情用户为翟天临进行辩解,正向观点多是发布在各类翟天临超话或话题等板块的广告类微博。而此时负向观点团簇数量较少,表明此时微博舆情用户关注点高度集中,主要对翟天临抄袭为娱乐圈与学术圈带来的影响、学术腐败问题以及官方发表声明与表态等内容进行讨论。对比两个关键时间节点中观点团簇的属性与团簇中观点内涵可以看出,“翟天临知网事件”中观点团簇的情感强度、传播范围与团簇规模都发生了极为明显的演化,并且负向观点团簇的语义内涵也从针对事件本身的讨论演化为该事件对社会秩序造成的影响。

(2)观点团簇演化链构建。在对该事件的网络舆情观点团簇演化属性进行测度之后,便可构建其观点团簇演化链,将前后两个时间节点中各个观点团簇进行交叉对比,将关键词相似度最高的前后两个时间节点中观点团簇关联起来便构成了观点团簇演化链。由于抓取的事件数据跨度时间较长,难以将全部网络舆情观点的演化链进行展示,因此在下文中以 2 月 9 日至 2 月 15 日的观点团簇数据展示其演化状态,经过网络舆情观点团簇演化链构建得到正向观点团簇演化链 23 条,负向观点团簇演化链 22 条。由于演化链较多,因此本次实验分别选取一条正向演化链与一条负向演化链的演化情况进行实证结果展示。正向观点团簇演化链与负向观点团簇演化链在 2 月 9 日与 2 月 15 日上前 10 个关键词变化对比如表 7 所示:

表 7　2 月 9 日与 2 月 15 日演化链前 10 个关键词

演化链极性	2 月 9 日	2 月 15 日
正向演化链	童年,人生,幼稚,心灵,奶奶,天真,烂漫,儿童,快乐,外婆	童年,人生,孩子,父母,母亲,快乐,生活,欢笑,幸福,美好
负向演化链	翟天临,事件,演员,学术,抄袭,北京大学,博士学位,论文,不端,造假	翟天临,事件,演员,学术,抄袭 造假,论文,北大,博士,北电

从表 6 可以看出,无论正向还是负向的演化链在这段时间内关键词语义内涵基本保持一致,表明这两条演化链都是各自种观点在时间上的延续,通过对演化链上观点团簇的演化属性数值进行对比,即可得到观点团簇的演化等级,如果数值增量较大则表明该演化链所代表的观点发生了明显的演化现象。通过测度后的正向观点团簇演化链与负向观点团簇演化链的演化属性与演化等级见表 8 与表 9。

表 6 与表 7 中团簇类别表示该演化链在每个时间段上由某个类别的观点团簇构成,该演化链中的观点团簇都表示同一种观点主题,从 10 日开始后的观点团簇都是 9 日起始观点团簇的延续,0 则表示该演化链没有后续演化,同时该演化链的演化属性也固定在某

表 8 正向观点团簇演化链的演化属性与演化等级

日期	团簇类别	传播范围	情感强度	团簇规模	演化等级
2 月 9 日	7	1 680	139	19	14 级
2 月 10 日	1	2 026	172	22	14 级
2 月 11 日	6	40 936	402	37	8 级
2 月 12 日	5	183 989	2 451	191	4 级
2 月 13 日	0	183 939	2 451	191	14 级
2 月 14 日	0	183 939	2 451	191	14 级
2 月 15 日	0	183 939	2 451	191	14 级

表 9 负向观点团簇演化链的演化属性与演化等级

日期	团簇类别	传播范围	情感强度	团簇规模	演化等级
2 月 9 日	4	893 032	-318	60	14 级
2 月 10 日	3	1 211 353	-761	124	14 级
2 月 11 日	5	15 539 456	-1 945	240	9 级
2 月 12 日	5	36 377 192	-9 617	819	6 级
2 月 13 日	3	37 879 618	-11 593	1 034	14 级
2 月 14 日	0	37 879 618	-11 593	1 034	14 级
2 月 15 日	0	37 879 618	-11 593	1 034	14 级

一数值上,3 种演化属性的数值则为该演化链上观点团簇的累加,从表 6 与表 7 可以看出,在 2 月 11 日与 12 日都处于较高的演化等级,观点团簇的情感强度、团簇规模和传播范围都有较大幅度的增长,说明 11 与 12 日的观点团簇有较为明显的演化现象,原因是 11 日北京电影学院宣布成立调查组对翟天临论文抄袭事件进行调查,引发 11 与 12 日连续两天的网民大规模讨论,形成网络舆情热度高峰,观点团簇演化等级在识别网络舆情演化高峰的辨识上具有较好的效果。

4.3 实证结果讨论

对于以"翟天临知网事件"为实证对象的网络舆情观点团簇演化等级测度结果表明,观点团簇演化等级测度模型能够提供各个观点团簇在传播范围、情感强度、群体规模、语义内涵等多个方面的准确数据,通过观测观点团簇的演化属性数值能够极快的掌握当前事件中影响能力最强的观点团簇,而对该团簇进行解构可以提取出团簇中的关键用户的信息,并实施具有针对性的观点引导策略。同时,根据进一步计算的观点团簇演化等级能够清晰地辨识观点团簇的演化状态,为网络舆情管控的快速反应与靶向引导提供依据。通过对实证结果的分析发现:①网络舆情观点团簇演化等级可应用于识别网络舆情中的群体观点在时间维度上的演化程度。例如,在本文案例"翟天临知网事件"中,在 12 日事件有新的进展并产生舆情热点,通过对该事件的观点团簇演化等级测度后得知在 12 日中演化等级最高,表明本文方法在识别观点团簇的演化上具有较好的效果。②网络舆情观点团簇的演化属性数值同样能够作为舆情引导的参考依据,从实证数据中可以看出,在该事件中呈现负向情感强度的观点团簇的各项属性数值皆远远大于正向情感的观点团簇,其中涉及到广大网友最为关注的问题的观点团簇的属性数值也远大于其他观点团簇,因此在实际舆情管控工作中可以根据观点团簇的属性数据进行有的放矢的靶向管控。同时,在某个时间点上的演化链数量也表示了该事件的观点群体数量,越多的演化链则群体观点越繁杂。③网络舆情观点团簇演化等级与演化属性数据可以清晰地表示目前网络舆情的演化态势,并且通过对照前序时间节点中观点团簇的各项数据能够得知处于同一演化链上的观点团簇的情感、规模、语义内涵、传播范围等信息的变化情况,提供多元化的数据支撑。

5 结语

笔者在理论层面对网络舆情观点团簇概念进行辨析,为网络舆情观点演化的测度提供新的思路。在实践层面,构建了网络舆情观点团簇演化等级测度模型,以"翟天临知网事件"为实证案例,对演化等级这一指标的准确性与可靠性进行验证,结果显示,网络舆情观点团簇演化等级能够精准的体现网络舆情群体思维的变化水平与演化程度,在网络舆情监管与预警上具有良好的适用性。

本文在研究中同样存在一定局限性:①在观点团簇的划分上基于 LDA 模型对网络舆情进行观点团簇划分,划分准确性有待提高;②本文主要对文本类型网络舆情的观点团簇演化进行辨别,缺少对图片、视频等多媒体内容演化状态的识别;③本文所提出的方法无法对观点团簇演化的未来趋势进行预测,仅能对当前时间节点上的演化等级进行测度。因此在下一阶段研究中,作者将对上述局限进行优化,在提高观点团簇划分准确率与适应多媒体网络舆情观点团簇演化感知的基础上,利用神经网络对观点团簇的未来演化趋势进行预测,以便于相关部门更好地管控网络舆情。

参考文献:

[1] 韩忠明,李梦琪,刘雯,等. 网络评论方面级观点挖掘方法研究综述[J]. 软件学报, 2018,29(2):417-441.
[2] 孙红,黎铨祺,赵娜. 基于双层树状支持向量机的观点挖掘与倾向分析[J]. 智能计算机与应用,2021,11(3):44-47.
[3] 唐洪婷,蔡秀定,张延林,等. 基于深度学习的企业开放社区用户创意挖掘方法研究[J/OL]. 系统工程理论与实践. [2021-09-21]. http://kns. cnki. net/kcms/detail/11. 2267. N.

20210527.1546.009.html.

[4] 张柳,王晰巍,黄博,等.基于字词向量的多尺度卷积神经网络微博评论的情感分类模型及实验研究[J].图书情报工作,2019,63(18):99 – 108.

[5] 陈健瑶,夏立新,刘星月.基于主题图谱的网络舆情特征演化及其可视化分析[J].情报科学,2021,39(5):75 – 84.

[6] 黄微,朱镇远,许烨婧,等.网络舆情衍进指数构建与实证分析[J].图书情报工作,2019,63(20):26 – 33.

[7] SAITO K, OHARA K, YAMAGISHI Y, et al. Learning diffusion probability based on node attributes in social Internets[C]//Foundations of intelligent systems-international symposium. Berlin: Springer berlin heidelberg, 2011.

[8] 周琦萍,杨芳.基于SIS模型的网络舆情无监督预警机制研究[J].情报科学,2019,37(8):51 – 55.

[9] 杨鼎,阳爱民.一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J].计算机应用研究,2010,27(10):3737 – 3739.

[10] BLEI D M, NG A Y, JORDAN M L, et al. Latent dirichlet allocation[J]. Journal of machine learning research, 2003, 11(3):993 – 1022.

[11] FREEMAN R M. Observing language changes in aging and Alzheimer's speech using information theory techniques[J]. Dissertations & theses - gradworks, 2015,11(1):21 – 26.

[12] 苏炯铭,刘宝宏,李琦,等.基于观点动力学的在线评分人数预测[J].计算机工程,2014,40(10):155 – 160.

[13] 王美方,刘培玉,朱振方.基于TFIDF的特征选择方法[J].计算机工程与设计,2007,28(23):5795 – 5796,5799.

[14] 高俊峰,黄微.网络舆情场中观点簇丛的情感极化度测算[J].图书情报工作,2019,63(10):106 – 114.

作者贡献说明：
闫璐：设计并修改研究方法，进行试验，分析数据，撰写论文；
杨刚：提出研究方向及思路；
赵江元：资料采集及整理，分析数据，文章结构修改。

Measurement and Empirical Study on the Evolution Level of Opinion Clusters of Internet Public Opinion

Yan Lu　Yang Gang　Zhao Jiangyuan
School of management, Jilin University, Changchun 130022

Abstract: [**Purpose/significance**] Proposing and constructing the evolution level of Internet public opinion clusters is to describe the evolution degree of the group opinion state of Internet public opinion audiences over time and events. It is of great theoretical and practical significance for Internet public opinion guidance and precise guidance. [**Method/process**] Based on LDA and CNN neural Internet, the paper constructed a level measurement model of Internet public opinion cluster evolution, and took "Zhai Tianlin CNKI event" as the experimental object to verify the effectiveness of the index of evolution level. [**Result/conclusion**] The evolution level of Internet public opinion cluster can well reflect the evolution of Internet hot event group opinion state. It can show the attribute values of three dimensions and also reflect the evolution degree of opinion clusters compared with the node state of the previous time. The evolution level measurement results of opinion cluster in this paper accurately reflect each evolution peak of event opinion. It provides a new direction for the relevant departments to target and guide the opinions of Internet public opinion groups.

Keywords: Internet public opinion　opinion cluster　opinion evolution　evolution level　measurement model